Buổi 2: Qui trình phát triển corpus

Đặc điểm của corpus

- "Chúng ta phải hiểu mình đang cần gì và xây dựng corpus đáp ứng yêu cầu đó."
- Machine-readable: Máy tính tính dễ dàng thực hiện các truy vấn, tính toán
  - Annotated data khó đọc hơn raw data.
- Reference to a standard
- Data đáng tin cậy, xảy ra một cách tự nhiên (natural occurring)

Đặc điểm của corpus

- Sampling:
  - Data được thu thập từ nhiều nguồn khác nhau để có thể làm mẫu được
  - data phải cân bằng
- Representativity: Có thể đại diện cho một ngôn ngữ hay nhiều ngôn ngữ
- Finite size
  - Một vài kho dữ liệu luôn tăng kích thước (them data mới): có thể update hiện tượng ngôn ngữ theo thời gian, …
  - Tuy nhiên:
    - Tính sampling and representativity khó đảm bảo
    - Tăng kích thước, tăng thời gian huấn luyện, …

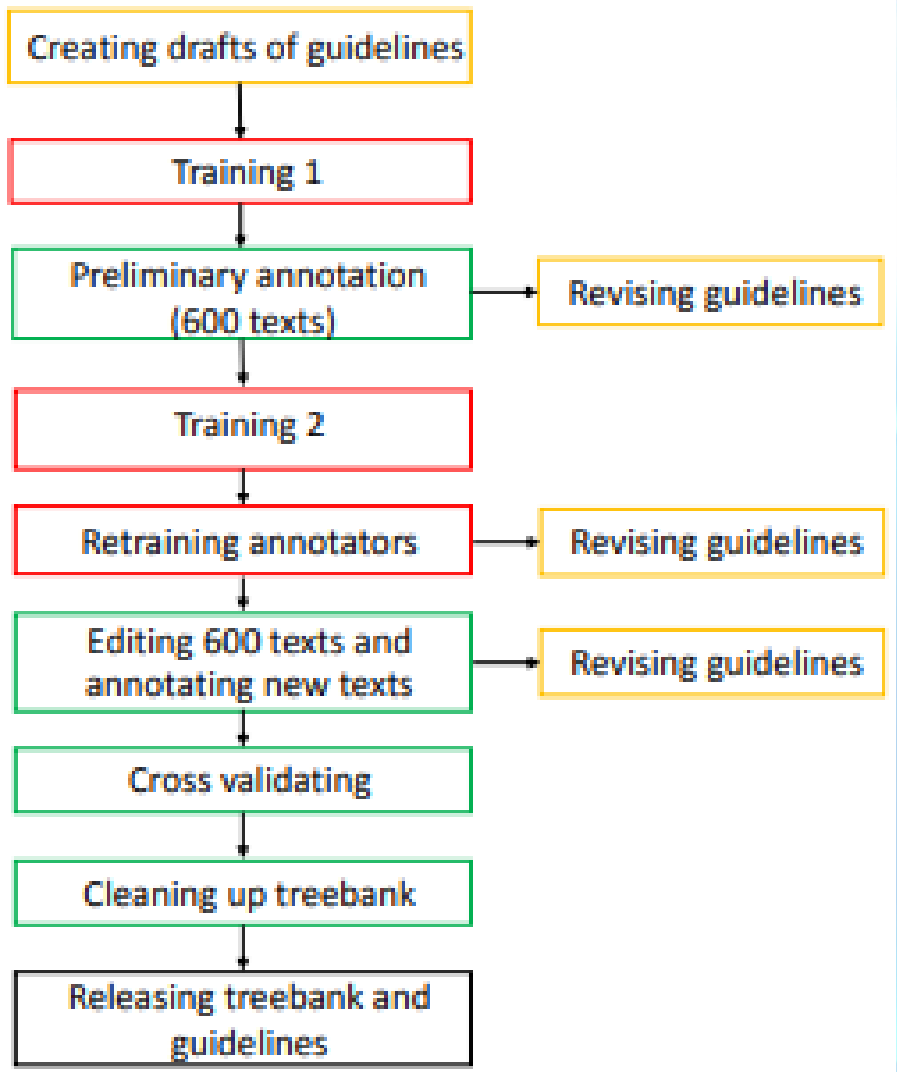
Qui trình phát triển corpus

- Thu thập raw text
- Phát triển guidelines
- Phát triển annotation tool và các tool tiền xử lý
- Huấn luyện annotator
- Markup & Annotation
- Kiểm tra chéo
- Cleaning up corpus
- Công bố corpus và guideline

Qui trình phát triển corpus

- Thu thập raw text
  - Naturally occurring
  - Sampled
  - Representative
  - Balanced

An appropriate annotation process



Mark-up

- Marking đặc trưng (features) của original text bằng cách thêm vào các mã (codes) để nhận biết các đặc trưng này
  - Layout ban đầu
  - Structure of the text: paragraph/sentence/chapter start/end points
  - Page breaks
  - Headings
  - Nguồn gốc của văn bản gốc (URL, sách, …)

⇒Mark-up can be used in corpus searches.

Ví dụ: Dựa vào heading code để tìm các heading xuất hiện trong text

Mark-up

- Marking extra information about the text (metadata)
  - Nơi xuất bản
  - Tên tác giả: Tác giả original text, tác giả tài liệu được annotate
  - Ngôn ngữ của tài liệu
  - Tập ký tự được sử dụng trong kho ngữ liệu
  - Điều kiện cấp phép
  - Nguồn gốc của văn bản gốc (URL, sách, …)

Metadata này có thể được để ở đầu mỗi corpus file

⇒Mark-up can be used in corpus searches.

Ví dụ: Lập danh mục các mục trong thư viện và các tài nguyên điện tử khác

Dublin Core - 15 Elements

- Nội dung (7)
  - Title, Subject, Description, Type, Source, Relation and Coverage
- Sở hữu trí tuệ (4)
  - Creator, Publisher, Contributor, Rights
- Khởi tạo (4)
  - Date, Language, Format, Identifier

International Standard Language Resource Number (ISLRN)

- Một nổ lực để cung cấp cho các tài nguyên ngôn ngữ một định danh duy nhất
- Giống như số ISBN được cấp cho sách
  - "Mục đích chính của lược đồ siêu dữ liệu (metadata scheme) được sử dụng trong ISLRN là để nhận biết nguồn tài nguyên ngôn ngữ. Lấy cảm hứng từ lược đồ được biết rộng rãi OLAC, một tập siêu dữ liệu tối thiểu được chọn để đảm bảo rằng bất kỳ nguồn tài nguyên nào cũng có thể được phân biệt và nhận biết một cách đúng..."

ISLRN Metadata schema

Metadata	Mô tả	Ví dụ
Title	Tên nguồn tài nguyên	1993-2007 United Nations Parallel Text
Full Official Name	Tên được sử dụng trong tài liệu tham khảo	1993-2007 United Nations Parallel Text
Resource Type	Bản chất hay thể loại của nội dung tài liệu theo quan điểm ngôn ngữ	Primary Text
Source/URL	URL, nơi chứa phiên bản đầy đủ của metadata	https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinSiteUEF
Format/MIME Type	File format của tài nguyên	Xml/text
Size/Duration	Kích thước tài nguyên	21416 KB
Access Medium	Phương tiện lưu trữ	Distribution: 3 DVDs
Description	Tóm tắt nội dung tài nguyên	
Version	Phiên bản hiện tại	1.0

ISLRN Metadata schema

Metadata	Mô tả	Ví dụ
Media Type	Danh sách các kiểu được sử dụng để phân loại bản chất hoặc thể loại nội dung tài nguyên.	Text
Language(s)	Ngữ liệu được viết hay nói bằng ngôn ngữ nào	eng (English)
Resource Creator	Người hay tổ chức chịu trách nhiệm chính ch việc xây dựng ngữ liệu	Ah Lian; NTU; lian@ntu; Singapore
Distributor	Người hay tổ chức chịu trách nhiệm chính ch việc công bố ngữ liệu	
Rights Holder	Người hoặc tổ chức sở hữu hoặc quản lý quyền đối với tài nguyên	
Relation	Các tài nguyên liên quan	

Annotation

Geoffrey Leech's Seven Maxims of Annotation

1. Có thể tách phần chú thích khỏi kho ngữ liệu đã được chú thích để trả lại dữ liệu thô
2. Có thể tách phần chú thích khỏi text.

   Kết hợp 1) và 2) => kho ngữ liệu nên cho phép thao tác linh động một cách tối đa
3. Lược đồ chú thích nên dựa trên guidelines được công bố cho người sử dụng (guidelines có thể có nhiều version)
4. Cần làm rõ việc chú thích dữ liệu được thực hiện như thế nào và bởi ai

Geoffrey Leech's Seven Maxims of Annotation

5. Người sử dụng dữ liệu nên biết rằng dữ liệu không phải là không có lỗi.
6. Lược đồ chú thích nên dựa trên các tiêu chí lý thuyết trung lập (không thiên vị) và được sự đồng thuận rộng rãi.
7. Không có lược đồ chú thích nào được coi là tiêu chuẩn. Tiêu chuẩn được thể hiện thông qua sự đồng thuận trong thực tế.

Types of Corpus Annotation

- Tokenization, Lemmatization (reduce inflectional forms)
- Parts-of-speech
- Syntactic analysis
- Semantic analysis
- Discourse and pragmatic analysis
- Phonetic, phonemic, prosodic annotation
- Error tagging

How is Corpus Annotation Done?

- Có 3 cách:
  1. Thủ công
  2. Bán tự động
  3. Tự động
- Đòi hỏi nhiều nhân lực: 1 > 2 > 3
- Vài annotation có thể làm tự động
  - Part-of-speech tagging: độ chính xác khoảng 97%
  - Lemmatization

  Tool gán nhãn tự động cho các loại nhãn khác chưa đủ tốt.

Lemmatization

- Các từ dẫn xuất có cùng nghĩa → Lemmatization: cắt giảm biến cách của từ (đưa về hình thức nguyên mẫu)
  - *unit* và *units*: 2 dạng từ khác nhau có cùng lemma là *unit*.
- Lemmatization áp dụng cho:
  - Hình thái số nhiều:
    - unit/units: unit
    - child/children: child
  - Hình thái động từ:
    - eat/eats/ate/eaten/eating: eat
  - Hình thái so sánh:
    - many/more/most: many
    - slow/slower/slowest: slow
    - much/more/most: much

Lemmatization

- Lemmatization không áp dụng cho:
  - Các từ dẫn xuất thuộc các nhóm từ loại khác nhau
    - quick/quicker/quickest: quick
    - quickly: quickly
    - Korea: Korea (noun)
    - Korean: Korean (adj)
- Lemmatization cho tiếng Anh có thể được thực hiện một cách đáng tin cậy bằng cách dùng tool tự động
- Tại sao cần phải Lemmatization?
  - Cải tiến information extraction tasks

Part-of-Speech (POS) Tagging

- (POS) Tagging: gán nhãn từ loại cho từ
  - Colorless/JJ green/JJ ideas/NNS sleep/VBP furiously/RB ./.
- Hữu ích: cùng một từ nhưng có nhãn từ loại khác nhau -> nghĩa khác nhau
- English Penn Treebank: 36 POS tags
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- Chinese Penn Treebank: 33 POS tags
  - http://languagelog.ldc.upenn.edu/myl/ctb-posguide.pdf

Penn Treebank Examples

Tag	Description	Tag	Description
NN	Noun, singular or mass	VB	Verb, base form
NNS	Noun, plural	VBD	Verb, past tense
NNP	Proper noun, singular	VBG	Verb, gerund or present participle
NNPS	Proper noun, plural	VBN	Verb, past participle
PRP	Personal pronoun	VBP	Verb, non-3rd person singular present
IN	Preposition	VBZ	Verb, 3rd person singular present
TO	to	.	Sentence Final punct (.,?,!)

- Nhãn từ loại có bao gồm thông tin biến cách
  - Biết từ loại, có thể tìm được từ nguyên mẫu
- Một vài nhãn rất specific
  - I/PRP wanted/VBD to/TO go/VB ./.

Universal Tagset

Tag	Giải thích	Ví dụ
VERB	verbs (all tenses and modes)	
NOUN	nouns (common and proper)	
PRON	pronouns	
ADJ	adjectives	
ADV	adverbs	
ADP	adpositions (prepositions and postpositions)	
CONJ	conjunctions	
DET	determiners	
NUM	cardinal numbers	
PRT	particles or other function words	
X	other: foreign words, typos, abbreviations	
.	. , ; ! punctuation	

Tìm từ trong tiếng Việt phù hợp với các nhãn trên?

Parsing (Syntactic Annotation)

- Parsing: thêm thông tin cấu trúc ngữ (parse) vào câu

  (S (NP (N Claudia))
  (VP (V ngồi)
  (PP (Cs trên)
  (NP (N ngưỡng_cửa)))))

- Parsed corpus còn được gọi là treebank
  - Penn Treebank, Chinese Treebank, Vietnamese Treebank, ...
- Công dụng: train automatic parsers
  - Stanford Parser và CMU parser được train trên Penn Treebank

Corpus vs. Annotation Software

- Chúng hỗ trợ lẫn nhau:
  1. Xây dựng ngữ liệu hoàn toàn thủ công
  2. Chương trình máy tính được huấn luyện sử dụng kho ngữ liệu này
  3. Sử dụng chương trình máy tính này để gán nhãn dữ liệu mới
- Trong thực tế, ...

Training a Parser/Learning a Model

- A corpus can be used to train a computer program. A program learns from corpus data. What does this mean?
  - *work* is a noun (NN) in some contexts, and a verb (VB) in some others.
  - When *work* follows an adjective (ADJ), it is likely to be a noun.
  - When *work* follows a plural noun (NNS), it is likely to be a verb.
    - nice/ADJ work/NN, beautiful/ADJ work/NN
    - they/NNS work/VB at a hospital
    - my parents/NNS work/VB too much

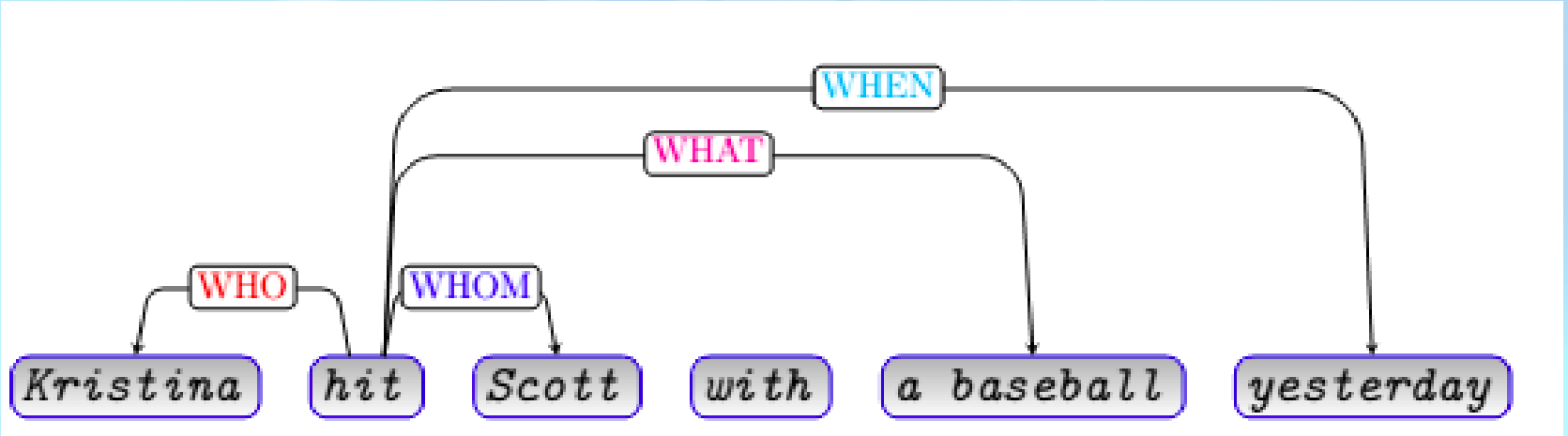
  These patterns can be extracted from a corpus, and the "trained" computer program makes a statistical model with them to predict the POS of work in a new text

Semantic Annotation

- Khử nhập nhằng ngữ nghĩa giữa các từ đồng âm
- Ví dụ:
  - từ *lie* trong:
    - *The boy lied$_1$ to his parents. (nói dối)*
    - *Mary lied$_2$ down for a nap. (nằm nghỉ)*
  - Từ *share* trong:
    - Mary did not share$_1$ her secret with anyone (chia sẻ)
    - The share$_2$ holders of Intel were disappointed (cổ phiếu)

Semantic Role Labeling

- Vai trò ngữ nghĩa là mối quan hệ giữa các thành phần cú pháp với vị ngữ (động từ chính):
  - Ví dụ: Agent (tác thể), Patient (bị thể), Instrument (công cụ), Locative (vị trí), Temporal (thời gian), Manner (cách thức), Cause (nguyên nhân), . . .



Semantic Role Labeling

- SRL rất hữu ích trong trong hỏi đáp. Nó giúp máy tính hiểu câu ở mức độ ngữ nghĩa nông, và có thể trả lời các câu hỏi sau đây:

Ai đã đánh Scott bằng một quả bóng chày?

Ai là người đã bị Kristina đánh bằng một quả bóng chày?

Kristina đã đánh Scott bằng cái gì?

Kristina đã đánh Scott bằng một quả bóng chày khi nào?

Time Annotation

- Biểu thức thời gian cho chúng ta biết:
  - Khi nào vấn đề xảy ra
  - Vấn đề xảy ra trong bao lâu
  - Vấn đề có xảy ra thường xuyên không
- Ví dụ
  - He wrapped up a three-hour meeting with the Iraqi president in Baghdad today.
  - The king lived 4,000 years ago.
  - I'm a creature of the 1960s, the days of free love.

Discourse and Pragmatic Annotation

- Chú thích các thông tin đàm thoại (thường áp dụng trong kho ngữ liệu bao gồm các đoạn hội thoại)
  - hành vi lời nói (ví dụ: chấp nhận, bày tỏ lòng biết ơn, trả lời, xác nhận, sửa, chào hỏi, …)
  - Các dạng hành vi lời nói (ví dụ: câu tường thuật, câu hỏi yes-no, câu mệnh lệnh, …)
- Coreference annotation:
  - theo dõi một thực thể được đề cập trong suốt văn bản
    - Kim$_4$ said ... Sandy$_6$ told him$_4$ that she$_6$ would ...
    - <COREF ID=’100’>The Kenya Wildlife Service</COREF> estimates that <COREF ID="101" TYPE=IDENT REF=’100’>it</COREF> loses $1.2 million a year in park entry fee...

Error Tagging

- Error tagging thường được làm cho các kho ngữ liệu dành cho người học
  - Cambridge Learner Corpora (CLC) and the Longman Learner's Corpus
- Các kiểu lỗi được sử dung trong CLC:
  - sử dụng sai từ
  - thiếu một cái gì đó
  - từ / cụm từ cần thay thế
  - từ / cụm từ không cần thiết
  - Ex: *my friend told me if I knew about Shakespare. But, <TIP id=17-56 etype=24 tutor="I knew">I know</TIP>*

Inter Annotator Agreement

- Q: làm thế nào để kiểm tra việc chú thích của chúng ta là đúng?
- A: Nhiều người chú thích cùng một dữ liệu, đo độ đồng thuận

Cohen's Kappa Coefficient

- Hệ số Kappa của Cohen: đo độ đồng thuận giữa 2 annotators

$$K = \frac{\mathrm{P}a - \mathrm{P}e}{1 - \mathrm{P}e}$$

  - Pa: độ đồng thuận giữa 2 annotators (accuracy)
  - Pe: xác xuất đồng thuận ngẫu nhiên (random agreement)

Cohen's Kappa Coefficient

Item #	1	2	3	4	5	6	7	8	9	10
Rater 1	R	R	R	R	R	R	R	R	R	S
Rater 2	S	R	R	O	R	R	R	R	O	S

	Rater 1				
Rater 2		R	S	O	Sum (row)
	R	6	0	0	6
	S	1	1	0	2
	O	2	0	0	2
	Sum (col)	9	1	0	

Tính Pa:

Pa = (6+1)/10 = 0.7

→ Rater 1 và rater 2 đồng thuận 70%

Tính Pe:

Xác suất **R1** chọn nhãn R: 9/10 = 0.9

S: 1/10 = 0.1

O: 0/10 = 0

Xác suất **R2** chọn nhãn R: 6/10 = 0.6

S: 2/10 = 0.2

O: 2/10 = 0.2

Xác suất **R1 và R2** chọn nhãn R: 0.9 * 0.6 = 0.54

S: 0.1 * 0.2 = 0.02

O: 0 * 0.2 = 0

=> Pe = 0.54 + 0.02 + 0 = 0.56

K = (0,7-0,56)/(1-0,56) = 0,3182

→ ***Quá thấp so với mức đồng thuận có thể chấp nhận (0,7)***

Tagging accuracy (machine)

Approaches to Annotation

- Nhiều annotator,
  - loại bỏ các ngoại lệ, hoặc
  - chọn các chú thích đa số annotator sử dụng
- 2 annotator, điều chỉnh các điểm không đồng thuận
- 1 annotator, điều chỉnh chú thích thông qua mô hình

How are corpora represented?

- Too many encoding schemes
  - TEI is common
  - Text
  - XML
  - XML standoff

Text Encoding Initiative

- Định nghĩa cách thức văn bản được mark-up bằng ngôn ngữ XML

XML

- XML (Extensible Markup Language): ngôn ngữ đánh dấu mở rộng
  - XML dùng để lưu trữ dữ liệu
  - Sử dụng các thẻ để đánh dấu văn bản
  - Các thẻ (tag) trong XML chưa xác định trước. Người dùng tự định nghĩa trong quá trình tạo tài liệu XML.
  - Các thẻ có thể được lồng vào đến độ sâu tùy ý

XML

```
<?xml version='1.0' encoding='UTF-8'?>
<article>
    <title>XML Cơ Bản</title>
    <author>Linh Nguyễn</author>
    <url>http://codehub.vn/xml-co-ban</url>
    <content>Bài viết giới thiệu về &gt;strong&lt;XML&gt;/strong&lt;...</content>
    <created_at>05/06/2017</created_at>
</article>
<article>
    <title>HTML Cơ Bản</title>
    <author>Ngọc Anh</author>
    <url>http://codehub.vn/html-co-ban</url>
    <content>Bài viết giới thiệu về &gt;strong&lt;HTML&gt;/strong&lt;...</content>
    <created_at>12/06/2017</created_at>
</article>
```

TEI Guildelines

- Mỗi text được markup theo TEI guidelines bao gồm 2 phần:
- Header (for markup)
  - Tác giả
  - Tựa đề bài báo
  - Ngày phiên bản, nhà xuất bản
  - …
- Body (for annotation)
  - Text được chú thích (gán nhãn)

Text

- One sentence per line, POS affixed

  can_VV

XML

- XML:

  <s sid=’1’><w wid = ’1’ pos=’vv’>can</w></s>

XML

<text>
<body>

<head>Subject: The staffing in the Commission of the European C
</head>
Can the Commission say:
1. how many temporary officials are working at the Commissio
2. who they are and what criteria were used in selecting the

<head>Answer given by <name type=PERSON><abbr rend=TAIL-SUPER>M
Cardoso e Cunha</name> on behalf of the Commission <date>(22 Se
1992)</date></head>
1 and 2. The Commission will send tables showing the number staff working for the Commission directly to the Honourable Mem
Parliament's Secretariat.
</body></text>

XML Stand off

- Tách riêng nhãn và text
- Ví dụ:
  - *text file:*

    This is a pen (text file)
  - corpus file:

    <pos=’noun’ cfrom=’0’ cto=’4’>

    <pos=’verb’ cfrom=’5’ cto=’7’>

    …

Maintained

- Errors corrected
- Cập nhật việc gán nhãn theo sự phát triển lý thuyết
  - Khi thay đổi mô hình, phải cập nhật chú thích

Case Study: the Hinoki Corpus

- Grammar-based syntactic annotation using discriminants
  - Parse the corpus and select the best parse
    - discriminant-based selection is efficient
  - Guarantees consistency
  - Loses some trees

Discriminant-based Treebanking

- Tính các biệt thức cơ bản (Carter 1997)
  - Sự trái ngược cơ bản giữa các phân tích
  - Hầu như độc lập và cục bộ
  - Có thể là cú pháp hay ngữ nghĩa
- Chọn hoặc loại các biệt thức cho đến khi còn lại 1 phân tích
  - $|\text{decisions}| \propto \log |\text{parses}|$
- Có thể loại tất cả các phân tích
  - ngữ pháp không thể phân tích thành công

How is Corpus Annotation Done?

- Mainly semi-automatic (done first by computer programs; post-edited)
  1. An small annotated corpus is built, entirely by humans
  2. Then a computer program is trained on this corpus
  3. Now new corpora can be automatically annotated using this program
- Large corpora often fully automatic
  - Segmentation
  - Part-of-speech tagging: accuracy of 97%
  - Lemmatization
- Corpora should indicate reliability of tags
  - Inter-annotator agreement, kappa (human)
  - Tagger accuracy (machine)

- http://www.geniaproject.org/genia-corpus/term-corpus
- https://www.clips.uantwerpen.be/conll2003/ner/
- http://2016.bionlp-st.org/tasks/bb2
- http://www.nactem.ac.uk/PHAEDRA/
- http://www.nactem.ac.uk/copious/copious_published.zip
- http://vlsp.org.vn/vlsp2018/eval/ner